

**METHOD AND SYSTEM FOR NAME-FACE/VOICE-ROLE
ASSOCIATION**

Inventor(s):

**Nevenka Dimitrova and Lalitha Agnihotri
Philips Research
345 Scarborough Road
Briarcliff Manor, NY 10510**

Assignee:

**Philips Electronics North America Corporation
1251 Avenue of the Americas
New York City, New York 10020**

METHOD AND SYSTEM FOR NAME-FACE/VOICE-ROLE ASSOCIATION

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to name and face/voice recognition systems. More particularly, the present invention relates to name-face-role association for consumer queries of databases.

2. Description of the Related Art

When we watch television and/or videos, sometimes we will notice a familiar actor whose name we may not be able to recall. Sometimes we can remember the name of the actor from a photo, but we can't recall the name of the character that he/she played in a particular movie. Sometimes, there is a desire to know what other roles a particular actor played, or possibly what are the names of the actors that have played a certain character over the years (for example, what are the names of all of the James Bond actors?)

In the prior art, there is a face-name link by CMU starts with visual features of the face and performs face recognition aided by closed captioning and text detection (superimposed text) if the name is below the face. The CMU system is meant for TV news programs. However, the face-name link falls short of a system accessible by consumers for providing name-face/voice-role association queries.

SUMMARY OF THE INVENTION

According to the present invention, there is a method and system for users to enter while viewing a television program, movie, sporting event, or by actually querying for name-face-role association to be output for further evaluation. There are many options available to a user after receiving the information, with one such option being to select a pay-per-view program, video rental or video download from the information returned by the database in response to the query. Another option would be to receive a video segment of the specific portion of program, movie, sporting event, etc, where a particular phrase has been spoken by a particular person, who may be an actor or athlete, or to provide a list of video segments.

Brief Description of the Drawings

Figs. 1A and 1B are a flowchart illustrating a name-to-face algorithm according to the present invention.

Fig. 2 illustrates examples of query boxes that can be used in a system for name-role-face association for consumer queries according to the present invention.

Fig. 3 illustrates one example of the arrangement of a system used for name-face-role association according to the present invention.

Detailed Description of the Invention

It is understood by persons of ordinary skill in the art that the following description, in conjunction with the accompanying drawings, is provided for purposes of illustration, and not limitation. A person of ordinary skill in the art understands that there can be variations in the depicted embodiments, which are within the spirit of the invention and the scope of the appended claims.

Figs. 1A and 1B comprise a flowchart illustrating an embodiment of a name-face-role recognition algorithm according to the present invention.

5 At step 100, there is an extraction of text boxes, followed by the application of text recognition. The extraction of text boxes may be performed a number of ways by using edge information, shape or color region segmentation. A role-name versus actor name list is generated from the recognized text.

At step 105 there is an extraction of face boxes, followed by the application of face recognition to the extracted faces. One possible method to extract face boxes is to use an eigenvector based method for face matching, including but not limited, the MIT Photobook, and the Virage Image system. In addition, model based face extraction is yet another way that the face extraction can be performed. In any of these methods, the result is that face models are generated.

Alternatively, or in addition to the extraction of face boxes, step 105 can include generating voice models by performing voice (speaker) identification using MFCC (Mel cepstral frequency coefficients). Voice identification is also well known in the art. Generating voice
20 models can greatly enhance the experience by the end user because an actor's voice is often his/her trademark. Even in cases where the actor mimics a certain type of speech, the actor's voice can be the most memorable part of a program.

At step 110, audio segmentation and speech to text conversion is formed if the closed captioning is not available. The audio segmentation and speech to text conversion is well known in the art.

At step 112, i is initialized to 1: $i=1$.

At step 115, the names from the beginning of a program (e.g. from the opening credits) or from the end of a program (from the ending credits) are extracted. Typically, these are main actors who carry the main roles. For example, at the beginning of the show, the names of the actors are often displayed on the screen, often superimposed on their picture, or a video segment. Typically, these names would be stored in a storage area of a database.

At step 120, a search for role-names in the transcript or closed captioning is performed. Closed captioning provides text of the people speaking during a video segment. A list is thus compiled of all the names found. This search is not limited to the role-names of the main actors and could be any recognized name. A first name on the list is referred in this embodiment as N_i .

Then, at step 125, the time code which accompanies the transcript may be used to search the corresponding video frames in the same, or even subsequent frames, for face models/voice models wherever N_i is found in the transcript.

At step 130, a fuzzy logic scheme is used to assign a degree of certainty C_i that one of the face/voices objects found in each of the corresponding video frames in step 125 is associated with the role name N_i . Of course, there are many instances where they may be several faces/voices in a video frame corresponding to or subsequent to a name in the transcript, and the

fuzzy logic provides one way to increase the degree of certainty. For each type of confirmation the degree of certainty is increased by a predefined increment. For example, if there is a mention of the name in the transcript the increment for the degree of certainty can be increased can be 0.1, if there is voice identity confirmation, the increase can be 0.2 and face identification can be 0.3.

Thus, the fuzzy logic ideally would initially assign an equal degree of certainty that the name found in the transcript is associated with one of the faces/voices found in the corresponding video frame or subsequent frame or frames.

At step 135, lip reading would be used to increase the degree of certainty by eliminating the face/voice of anyone that says the name identified in the transcript. The reason is that people normally do not say their own name. Of course, there are instances in a show where a person might be asked to say their name, but that would only serve to potentially reduce the degree of certainty, as a person would not normally continue to pronounce their own name again and again.

At step 140, the rest of the transcript is scanned and certainties for each subsequent appearance of a role-name is updated.

At step 145, there is a decision, where it is determined whether the certainty of a face model/voice model with a role-name association N_i is above a predetermined threshold. If the threshold has not been reached, the process goes back to step 115 and the steps up to 145 are

performed again until the threshold is reached. It is understood by persons of ordinary skill in the art that with each pass the fuzzy logic will improve the certainty of the face model until that threshold is reached.

5 At step 150, when the threshold has been reached for face model N_i , there is backpropagating to mark all the segments where the face model appeared for future identification.

At step 155, the role-name and actor name, and the face models/voice models are stored in a database for future retrieval.

At step 160, there is a decision as to whether N_i is the last element of the role-name list generated at step 100. If N_i is not the last element of the list, the process goes back to step 115 to repeat the steps for the next name, N_{i+1} . If N_i is the last element of the list, the process continues at step 165.

Finally, at step 165, if there are no more elements on the list, the face models/voice models N_i to N_{i+n} are sorted to keep the models having the highest degree of certainty according to a predetermined threshold in a database. It should be understood by persons of
20 ordinary skill in the art that the items stored in the database as a role appearance data structure, where the role name, duration time in the video segments, pointers to face models/voice models and the degree of certainty of a correlation of the actor name-role name with the face

model/voice model is stored. These items may be updated each time the same role is mentioned, and the degree of certainty can be increased or decreased.

Figure 2 illustrates query boxes for a system using name-role-face/voice association for consumer queries according to the present invention. Item 205 represents a database. The database contains, for example, for each movie: a role-name with an actor name list; for each role name, there are the video segments that the role name appears in; for each video segment, there are sentences spoken by the role. This database 205, can be separate from, or in communication with a database 210 of directors/movies/actors. The aforementioned process can create/update the database 205 based on information contained in the database 210.

The consumer queries 215, 220, can be in communication with the database over the Internet, fiber optic network, RF network, LAN, WAN , just to name a few possible forms of communication.

The consumer may present queries in a number of varied ways. For example, if the consumer is interested in the movie *Casablanca*, he/she may ask who acts in the role of Rick Blaine? The database 205 would be searched and return the name of Humphrey Bogart because of the role-name versus actor name list. As well as provide access to all the movie scenes (segments) with Humphrey Bogart in the role of Rick Blaine. In addition, the consumer could receive a photo of Humphrey Bogart from a face model from the movie. Further, the consumer could receive audio of Humphrey Bogart speaking, or in lieu of a photo, the consumer could

receive a video segment with audio of Humphrey Bogart speaking. The video segment could be any format, including, but not by any means limited to, MPEG.

The consumer could also query as to when did Rick Blaine first appear in the movie?

5 In the movie, Rick Blaine is first seen playing chess by himself inside Rick's Cafe, several minutes into the movie. This exact video segment can be retrieved from the database and sent to the consumer. In addition, a summary of the movie from Rick Blaine's point of view could be provided including video segments, combination of snapshots and part of the transcript.

Another example, if the consumer knew that an actor once said in a movie "Here's looking at you kid" there could be a query where the consumer asks who said that phrase, and what movie is it from. The database would then identify that the phrase was said by the role-name Rick Blaine in the movie *Casablanca*. The actor name of Humphrey Bogart and a face model may be retrieved and provided to the consumer. Alternatively, a segment of Humphrey Bogart saying the phrase "Here's looking at you kid" can be provided.

Additionally, the consumer may query what role that Ingrid Bergman plays in the movie *Casablanca*. The database would retrieve the role-name of Ilsa Lund. It could be queried as to what she said in the movie. Video segments for each of the times that Ilsa spoke could be
20 retrieved. Alternatively, text and/or audio could be retrieved where she said "Play it Sam. Play 'As Time Goes By.'" In turn, the image and/or voice of Sam, and his actor name Dooley Wilson, may also be retrieved.

is preferred to be an eigenvector based method such as, for example, MIT Photobook and/or Virage Image System be used for face models, and in the case of voice (speaker) identification, the use of, for example MFCC (Mel frequency cepstral coefficients). The mel-scale is a frequency-binning method which is based on the ear's frequency resolution, because the ear does not perceive sounds linearly across the audio spectrum. By the use of frequency bins on the mel-scale, mel-cepstral coefficients (MFCC) are computed so as to parameterize speech. The MFCC is a good indicator of the discrimination of the ear. Accordingly, MFCC can be used to compensate distortion channels through implementation of equalization by subtraction in a cepstral domain, as opposed to multiplication in a spectral domain.

Voice identification is also well known in the art. The faces/audio (and the detection of text) would be extracted from a video segment of program 318. This of course, can be in any format compatible with the extraction module. It should be noted that if the detection module 320 does not detect closed captioned text, then a speech-to-text conversion module (not shown) would be required.

The lip reading module 330 would be used for reading lips of the actors to identify whether the actors say the role-name. The more the actor says the role-name, the less likely is that particular actor to be the person playing that role-name, because it is unusual for people to keep repeating their own name.

The user display 335, could be a computer terminal, hand held PALM, Pronto remote control, display of a personal computer, or even a television set. While the system can be

Internet based (or other types of networks) portions of the system could be arranged as part of a television display or advanced VCR/DVD player, advanced set-top box, which permits querying capability. It would be preferable for the storage database to be remotely situated because of size constraints and capacity.

5

It is also envisioned that not only could names, sentences, video segments, or a list of video segments including audio be output, but the consumer would have the option of, for example, purchasing, renting or downloading the movie from which the video segment/segments are provided in response to their query. Accordingly, while the consumer may be using a computer to make the query, an intelligent television system, which permits queries by keying in data, or even voice commands, could be used.

The following is a description of one way that the present invention could be used in a home video. For example, very often family videos are a long (sometimes hours long) compilation of activities. In a home video, there would be roles, such as mom, dad, grandma, grandpa, baby, uncle, aunt, cousin, etc. In this particular case, instead of using transcripts as in the case of movies/programs, a user can assign the above-named roles (as well as many other roles in the video). After the roles are assigned, the voices can be determined, particularly if the assigned roles speak in the segments used to assign the roles. Thus, for example, in a videotape of a family reunion, or wedding, the ability to catalog all of the guests, and later retrieve their image and/or voices wishing congratulations can be accomplished by the presently claimed invention as unknown heretofore.

In another variation, instead of the roles of family members, such as a recorded meeting or a video conference, there could be assignments such as "boss," "the CEO," "notes taker," "client," "sales representative", etc. Alternatively, in a legal teleconference, the judge, plaintiff, defendant, plaintiff's attorney, defendant's attorney, jury, bailiff, and witnesses could also be identified by the presently claimed invention. If the meeting is in a courtroom, the transcript could also be an item used to assist in identification of the parties.